

Supporting Information

Degnan et al. 10.1073/pnas.0900194106

SI Methods

DNA Isolation and Construction of Libraries. Two complementary sequencing strategies were required to finish the *H. defensa* genome: (i) subcloning and Sanger sequencing a large insert BAC library and (ii) pyrosequencing (454 Life Sciences/Roche Applied Sciences). Live *A. pisum* str. 5AT aphids (9.15 g) infected with *H. defensa* and the phage APSE-2 were homogenized in PBS, serially filtered, and centrifuged to isolate the bacterial cells as described previously to minimize contaminating DNA (1). For the BAC library, purified cells were imbedded in 0.75% PFGE-certified agarose (BioRad), lysed, and then stored in low TE buffer (10 mM Tris-HCL, 0.1 mM EDTA pH 8.0) at 4 °C. The DNA containing plugs were restricted with HindIII at 37 °C for 20 min and DNA was separated on a pulsed-field gel. DNA ranging from 100 to 300 kb was extracted, purified, and ligated into pAGIBAC1 vector (2).

BAC clones were sequenced bi-directionally and binned by Blast identity (gamma-proteobacteria, *Buchnera*, Bacteria, insect) and %G + C. Putative *H. defensa* clones were fingerprinted, assembled into scaffolds with fingerprint contigs, and minimal tiling paths were chosen (2). Individual BACs were then subcloned into pUC19 vector, sequenced bi-directionally on an ABI3730xl, and assembled using Phred, Phrap, and Consed (3–5). Overlapping and validated BACs were then merged.

Pyrosequencing was performed to complete the nonclonable regions of the chromosome. *H. defensa* DNA was purified for 2 separate single-stranded template DNA (sstDNA) libraries in a similar fashion as above. However, instead of embedding the bacterial cells in agarose, they were resuspended in 500 μ l PBS with 2 U of rDNase I and 55 μ l of DNA-free 10 \times buffer (Ambion) and incubated for 1 h. at 4 °C. The rDNase I was inactivated by adding 0.5 M EDTA to a final concentration of 50 mM. The cells were then pelleted and washed twice with PBS, centrifuging at 4,000 \times g for 10 min at 4 °C each time. High molecular-weight DNA was then isolated from the pelleted cells using the Puregene Tissue Core Kit B (Qiagen). The manufacturer's instructions were followed for gram-negative bacteria, except for an additional treatment of the sample with 3 mAU Proteinase K for 30 min at 55 °C before the RNase A treatment.

We generated a standard and paired-end sstDNA library using the GS DNA Library Preparation Kits (Roche Applied Sciences), starting with 8 μ g and 15 μ g of purified DNA, respectively. The sstDNA libraries were then amplified by emPCR and sequenced on a GS-FLX (454 Life Sciences). A half run was used for the standard library and a quarter run for the paired-end library. The 454 reads were assembled with Newbler (v1.1.03.24) using default parameters. The resultant contigs were screened and sorted based on read coverage, G + C% and Blast.

Genome Closure. Completion of 2 of the 3 nonclonable regions was straightforward given the contigs generated by Newbler. The third region was highly repetitive, corresponding to one of the integrated copies of pHD5AT. Scaffolds of this region were confirmed by sequencing through overlapping PCR products (performed as in ref. 6). When necessary, PCR products were cloned into the TOPO-TA vector (Invitrogen). The order, orientation, and merger of scaffolds were determined by combinatorial PCR between scaffold ends. Sanger reads from PCR products and clones were merged with the 454 contigs in Consed using Phrap.

Phylogenetic Analyses. Lerat et al. (7) identified 203 single copy orthologs (SICO) in 13 gamma-proteobacterial genomes and used them to generate a species phylogeny. We used the *E. coli* copies of these genes and a BlastP bit score ratio threshold (>0.3) to identify orthologs in *H. defensa* and 28 additional genomes [supporting information (SI) Table S6]. Bit-score ratios were calculated based on the bit score of hit divided by the maximal bit score for the query sequence. Protein sequences of the identified SICO genes present in all 30 genomes were aligned in Muscle v3.6 (8), and invariant- and gap-containing columns were removed. Individual protein alignments were then concatenated into 1 of 4 alignments: all proteins; proteins involved in transcription, translation, or replication; *H. defensa* proteins with divergence from *E. coli* of <0.2; and *H. defensa* proteins with $\geq 44\%$ G + C. Pairwise protein divergences were estimated in PAML (9). Alignments with the *H. defensa* and *R. insecticola* sequences removed were generated to assess the possibility of long-branch attraction or other artifacts. Each dataset was then analyzed with RaXML and PhyML (10, 11). The best topologies were estimated using a gamma model of rate heterogeneity and either the WAG (RaXML) or JTT (PhyML) model of amino acid substitution. Unique topologies were then analyzed using the SH-test in TREE-PUZZLE 5.2 (12). The topology with the lowest log L and that disagreed with the fewest datasets is presented, and reported support values were estimated from 100 nonparametric bootstrap replicates.

***H. defensa* Proteomic Analysis.** Intact *H. defensa* str. 5AT cells were isolated as above from infected pea aphids in 2 separate batches, starting from 1.0 g and 4.5 g of mixed-age individuals. After filtration and centrifugation, the bacterial cells (with some insect and *Buchnera* contamination) were pelleted and frozen at –80 °C. These samples were processed and run using a SCX-RP LC-MS/MS device (MudPit analysis) following the standard protocols for the University of Albany Proteomics Facility. Briefly, the sample was homogenized in a buffer containing 2% SDS, 100 mM Tris, 10 mM DTT at pH 7.5, then centrifuged at 100,000 \times g. The supernatant was recovered and precipitated with 10% TCA then washed 3 times in cold acetone. The protein pellet was dissolved in a SDS loading buffer and loaded on a 10% SDS/PAGE gel. After running, the entire lane was cut into 12 sections, each of which was alkylated and subjected to in-gel tryptic digestion.

The tryptic peptides were extracted from the gel, concentrated then injected into the LC-MS/MS system, which comprised a Q-TOF 2 mass spectrometer (Micromass) equipped with the CapLC system (Waters). A Magic C18 LC column was used [5 μ m C18 particles, 100 μ m ID \times 150 mm (Michrom Biore-sources)]. Mobile phase A consisted of 0.1% formic acid, 5% acetonitrile, and 0.01% TFA, and mobile phase B consisted of 0.1% formic acid and 0.01% TFA in an 85/10/5 acetonitrile/isopropanol/water solution. The peptides were eluted during 78 min under a linear gradient from 15 to 50% mobile phase B with a final flow rate of 250 nl/min.

The Mascot generic-format peak list was created using Mascot distiller 2.0 software (Matrix Science). Mascot 2.2 search engine was then used to assist the interpretation of tandem mass spectra against a custom sequence database. This database comprised 2,155 predicted *H. defensa* CDS (CP001277, CP001278), 565 predicted *B. aphidicola* CDS (NC_011833, NC_002252, NC_002253), and 10,499 *A. pisum* proteins from RefSeq. The following search parameters were used: trypsin-specificity re-

striction with 1 missing cleavage site, variable modifications including deamidation (N,Q), oxidation (M), carbamidomethylation (C), and peptide and fragment mass tolerances were set to 0.3 Da. Peptide matches with Mascot significance scores <0.05 , and Mowse ion scores >31 were considered. Using a decoy database of the reverse of the custom protein database, the false-discovery rate was estimated to be 0.17% for the *H. defensa* peptides (3 *H. defensa* peptides found in the decoy database vs. 1,748 in the real database).

SI Results

H. defensa str. 5AT possess a 2,110,331-bp circular chromosome and a 59,034 bp conjugative plasmid. The chromosome is significantly larger than our previous estimates based on pulsed-field gel electrophoresis (≈ 1.7 Mb) (1), a disparity that likely arose for the limits of our resolving power on the pulsed-field gels and restriction fragments of similar sizes. Despite the abundance of putative *H. defensa* BACs in the library (913 out of 1,536), a single tiling path was not identified because of nonclonable regions (e.g., origin of replication, APSE prophage) (Fig. S2). The sequence gaps between the 3 supercontigs were closed using 454 pyrosequencing reads and confirmed with Sanger sequences of PCR amplicons. The two 454 pyrosequencing runs generated 330,169 reads, totaling 70,556,348 nt and averaging 213.7 nt per read. The vast majority was found to be from *H. defensa* (306,801). Complications during the final assembly process lead to the identification of the conjugative plasmid pHD5AT. Without previous evidence of *H. defensa* str. 5AT possessing an extrachromosomal element, it would have been overlooked.

The chromosome and plasmid assemblies were finished to accepted standards; however, a single uncertain region of the chromosome remains (nt 1,351,906–1,352,006). This region is associated with a large (≈ 70 kb) inverted, segmental duplication. Although this duplication was supported by several BACs, we were unable to determine the exact terminus of the duplication by long-range PCR or inverse PCR. However, given the even genome coverage of the 454 sequences (28.7X), and the absence of any unplaced *H. defensa* contigs, we do not believe that we are missing any gene content in the current assembly.

Metabolism of *H. defensa*. Carbon, in the form of the hexose sugars mannose, fructose, glucose, and glucosamine, is acquired by *H. defensa* from the aphid through one of three phosphoenolpyruvate-dependent, sugar transporting phosphotransferase systems. The latter 3 are abundant in sap-feeding insects, as fructose and glucose constitute sucrose, which is the major component of most plant saps (13) and glucosamine is a principal constituent of insect exoskeletons. The transported sugars are catabolized during glycolysis, the pentose-phosphate pathway, and the TCA cycle, forming ATP, reductants, and essential precursor metabolites. Oxygen is used as the terminal electron acceptor by the cytochrome *bo* oxidase complex, generating the proton-motive force that drives subsequent cellular processes. However, under oxygen limitation, *H. defensa* can also ferment pyruvate and acetyl-CoA to generate NAD^+ and ATP.

The reduced genome of *H. defensa* exhibits little redundancy in pathways for the biosynthesis or transport of amino acids, vitamins, and cofactors. For example, it can only synthesize 2 essential amino acids but has active transporters for the remaining 8 (Table S1). Within aphids, the obligate nutritional mutualist *Buchnera* synthesizes these essential amino acids, from which *H. defensa* clearly benefits. Tsetse flies also host both an obligate endosymbiont *Wigglesworthia glossinidia* that provides the host with B vitamins absent in vertebrate blood, and the endosymbiont *Sodalis glossinidius*, which has no known benefit for the tsetse fly (14, 15). In contrast to *H. defensa*, *S. glossinidius* exhibits minimal metabolic reliance upon the host or *Wiggles-*

worthia, as genes are present for the synthesis of all 20 amino acids and all of the vitamins and cofactors, save thiamine (see Table S1) (15). The *Sodalis* genome also has extensive evidence of recent gene inactivation, resulting in a coding density of only 51%. Together, these data are interpreted as evidence of *Sodalis* having undergone a recent shift from a free-living bacterium to a symbiont of tsetse flies (15, 16).

Putative Virulence Mechanisms. Nearly 6% of the *H. defensa* CDS are homologous to putative virulence factors (123 out of 2,100). This includes two type-3 secretion systems (T3SS) similar to SPI-1 and SPI-2 of *Salmonella typhimurium* LT2. We identified 10 CDS with similarity to known secreted effector proteins in other Bacteria (Table S2). Two effectors remain within the SPI-2-like genomic island, *spiC* (*ssaB*) and *ssaE*. Another 7 putative effector proteins are distributed throughout the genome and were likely horizontally transferred from disparate sources. The effectors include a tyrosine protein phosphatase (*yopH*), cysteine protease (*yopT*), and 2 copies of an ADP-ribosylating toxin (*aexT*) and adenylate cyclase (*exoY*). The N-terminal region of the last putative effector is similar to the non-LEE encoded effector D (*nleD*) from enterohemorrhagic *E. coli*, although the *H. defensa* CDS is 4-times as long (232 AA vs. 1,053 AA).

Additional virulence factors in the *H. defensa* genome include a potentially inactivated copy of a cytotoxic necrotizing factor-like protein (HDEF_2319, HDEF_2320) and 2 copies of the membrane-associated protein *mviN* (HDEF_0205, HDEF_0640). *H. defensa* also carries multiple classes of putative adhesins that may be involved in cell recognition, adherence, or invasion. In particular, several ORFs have significant similarity to type-5 secretion systems (T5SS), containing autotransporter barrel domains and pertactin adhesin domains. Additionally, *H. defensa* encodes a tight adherence locus (*tad*) originally identified in *Actinobacillus actinomycetemcomitans*, but widespread among Bacteria (17), and 3 chromosomally encoded pili clusters. However, most of the chromosomal pili gene clusters have several genes that are missing or inactivated.

***H. defensa* Proteomics.** A total of 11,491 tandem mass spectra were detected during the MudPit analysis of the *H. defensa* str. 5AT protein sample. Using stringent significance cutoffs (<0.05 Mascot score, >31 Mowse ion score), 5,984 peptides (52%) were identified as matching proteins from *H. defensa*, *B. aphidicola*, or *A. pisum*. Of these, 1,748 peptides (29%) were from *H. defensa* and corresponded to 89 proteins (see Table S5). Lowering the significance thresholds (<0.1 Mascot score, >0 Mowse ion score) increased the overall number of peptides matching the database ($n = 11,166$), but only identified 5 additional *H. defensa* genes (*ClpX*, *YfgL*, *RplX*, *RpsI*, *Tig*). The data reported hereafter only include the conservative peptide matches.

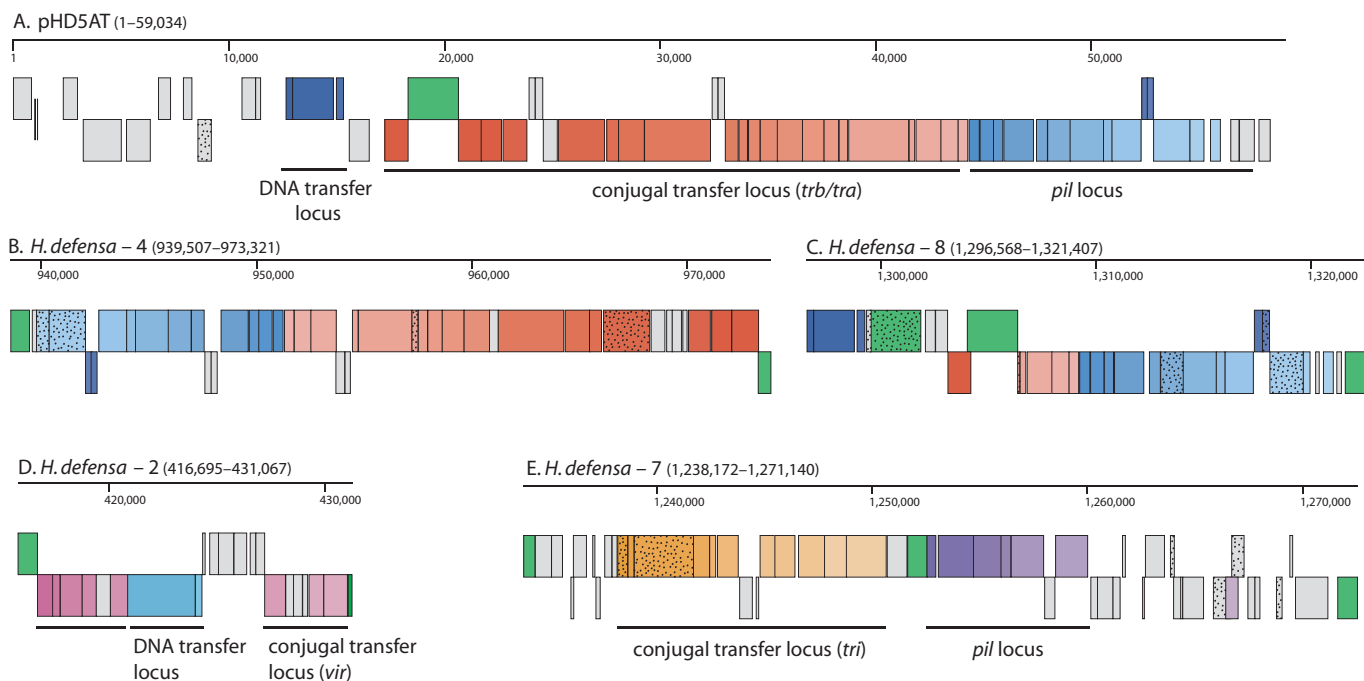
Categorizing the expressed proteins by functional roles revealed that core cellular processes, such as transcription, translation, and glycolysis were well represented (34 out of 89). Proteins involved in stress response (chaperonins, proteases) and various membrane components (transporters, lipoproteins) were also abundant. Additionally, 5 APSE proteins were identified, including the major capsid head protein (P24), suggesting the *H. defensa* cell isolation captured intact or precursors of phage particles. Also of note were 3 peptides that matched SseC, a translocon component of the SPI-2-like T3SS.

The proteins were ranked by relative abundance using the exponentially modified protein abundance index (emPAI) (18). This estimator, while crude, quantifies protein abundances by considering the number of observed peptides and the number of observable peptides per protein. Twelve *H. defensa* proteins had emPAI values >1 , suggesting elevated expression (Fig. 4). Among these are 4 genes involved in stress responses and 4 genes involved in membrane structure or transport. The elevated

expression of these 2 categories mirrors what has been observed in the *Buchnera* (19). In fact, the chaperone Hsp60 (GroEL, MopA), the most highly expressed *H. defensa* gene, is 40-fold higher than the majority of proteins detected ($n = 77$). This is also true of *Buchnera*, where GroEL is constitutively expressed and represents 10% of the proteins it synthesizes (20). Chap-

eronin is also highly expressed in a number of other obligate and facultative insect endosymbionts [*Portiera* (whiteflies), SOPE (weevils), *S. glossinidius* and *W. glossinidia* (tsetse flies)] (21–23). A possible explanation for this pattern is that increased GroEL expression buffers the genome-wide fixation of slightly deleterious alleles because of genetic drift (24).

1. Moran NA, Degnan PH, Santos SR, Dunbar HE, Ochman H (2005) The players in a mutualistic symbiosis: insects, bacteria, viruses, and virulence genes. *Proc Natl Acad Sci USA* 102:16919–16926.
2. Peterson DG, Tomkins JP, Frisch DA, Wing RA, Paterson AH (2000) Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide. *J Agr Genomics* 5.
3. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194.
4. Ewing B, Hillier L, Wendel MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175–185.
5. Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195–202.
6. Degnan PH, Moran NA (2008) Diverse-phage encoded toxins in a protective insect endosymbiont. *Appl Environ Microbiol* 74:6782–6791.
7. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the γ -Proteobacteria. *PLoS Biol* 1:e19.
8. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
9. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13:555–556.
10. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
11. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
12. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
13. Sandström JP, Pettersson J (1994) Amino acid composition of phloem sap and the relation to intraspecific variation in pea aphid (*Acyrtosiphon pisum*) performance. *J Insect Physiol* 40:947–955.
14. Akman L, et al. (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat Genet* 32:402–407.
15. Toh H, et al. (2006) Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. *Genome Res* 16:149–156.
16. Darby AC, et al. (2005) Extrachromosomal DNA of the symbiont *Sodalis glossinidius*. *J Bacteriol* 187:5003–5007.
17. Kachlany SC, et al. (2000) Nonspecific adherence by *Actinobacillus actinomycetem-comtans* requires genes widespread in Bacteria and Archaea. *J Bacteriol* 182:6169–6176.
18. Ishihama Y, et al. (2005) Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Mol Cell Proteomics* 4:1265–1272.
19. Maezawa K, et al. (2006) Hundreds of flagellar basal bodies cover the cell surface of the endosymbiotic bacterium *Buchnera aphidicola* sp. strain APS. *J Bacteriol* 188:6539–6543.
20. Baumann P, Baumann L, Clark MA (1996) Levels of *Buchnera aphidicola* chaperonin GroEL during growth of the aphid *Schizaphis graminum*. *Curr Microbiol* 32:279–285.
21. Aksoy S (1995) Molecular analysis of the endosymbionts of tsetseflies: 16S rDNA locus and over-expression of a chaperonin. *Insect Mol Biol* 4:23–29.
22. Charles H, Heddi A, Guillaud J, Nardon C, Nardon P (1997) A molecular aspect of symbiotic interactions between the weevil *Sitophilus oryzae* and its endosymbiotic bacteria: over-expression of a chaperonin. *Biochem Biophys Res Commun* 239:769–774.
23. Salvucci ME, Stecher DS, Henneberry TJ (2000) Heat shock proteins in whiteflies, an insect that accumulates sorbitol in response to heat stress. *J Therm Biol* 25:363–371.
24. Fares MA, Ruiz-Gonzalez MX, Moya A, Elena S, Barrio E (2002) Endosymbiotic bacteria: GroEL buffers against deleterious mutations. *Nature* 417:398.



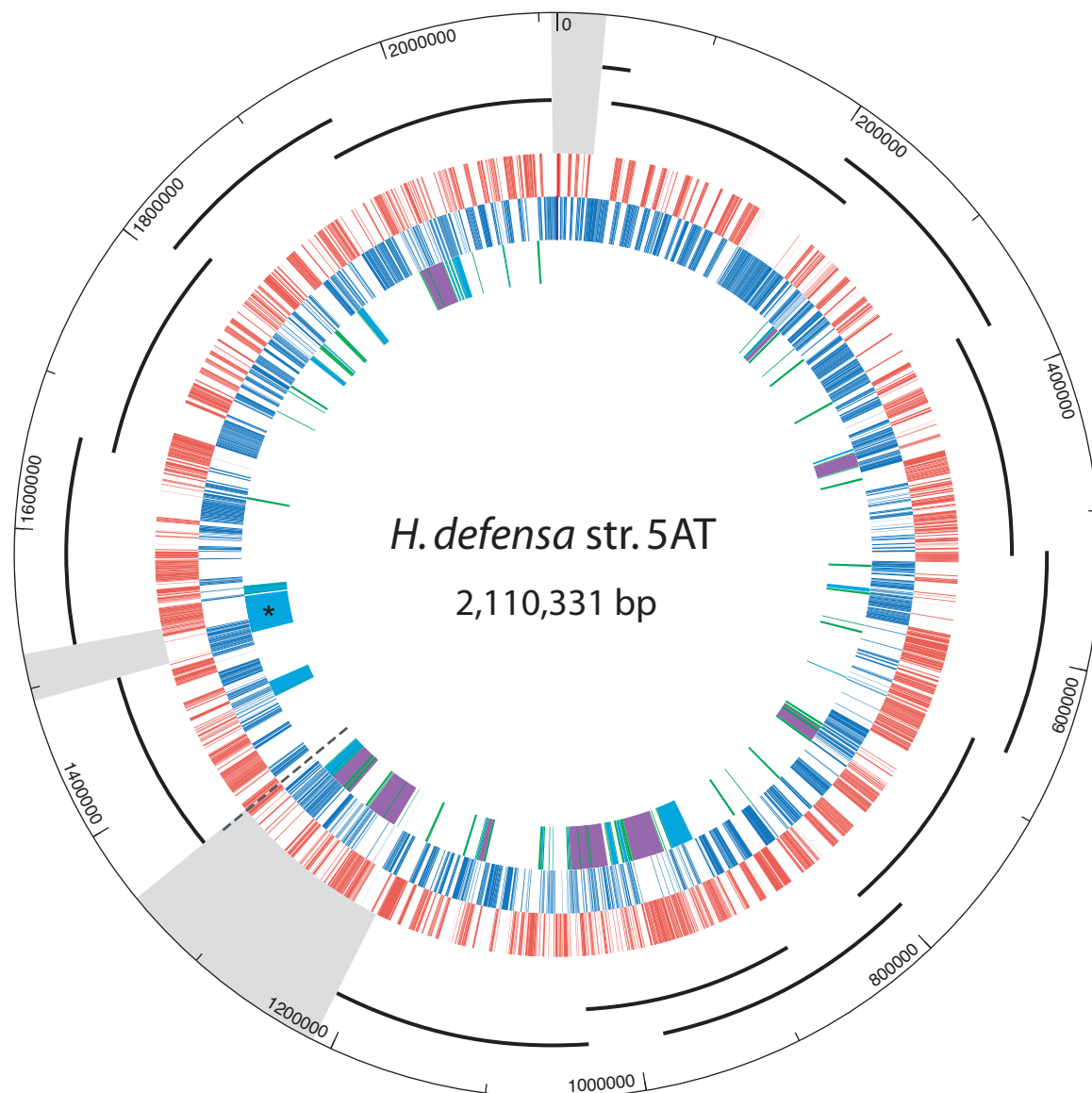


Fig. S2. BAC coverage of *H. defensa* genome. Starting from the outer ring: (i) coordinates (bp), (ii) location of overlapping BAC clones (solid black lines) and nonclonable regions (gray regions) closed by pyrosequencing, (iii) CDS on the forward strand (in red), (iv) CDS on the reverse strand (in blue), and (v) mobile genetic elements, IS elements, and group II introns (green), phage (blue), or plasmid (purple) islands. An asterisk indicates the location of the APSE prophage and the dashed line in (iii) to (v) is the location of the incomplete genome juncture.

Other Supporting Information Files

[Table S1](#)
[Table S2](#)
[Table S3](#)
[Table S4](#)
[Table S5](#)
[Table S6](#)
[Table S7](#)